图书情报工作
第65卷 第21期 2021年11月

融合篇章结构的文本知识网络构建*

■ 刘耀[1] 张越[2] 叶璐[3]
[1] 中国科学技术信息研究所 北京 100038 [2] 密歇根州立大学 东兰辛 489132
[3] 北京大学软件与微电子学院 北京 100871

摘 要：[目的/意义]文本向量化处理是文本挖掘、信息检索、情感分析等领域必须要经过的预处理过程，使节点向量包含丰富且有效的语义及结构信息是目前亟待解决的问题。[方法/过程]首先对科技政策类的文本特征进行分析，分别依照概念与概念间关系的分类体系，用 BiLSTM-CRF 算法和 SVM 分别实现对概念与概念关系进行自动标引，在特征工程同时融入基本特征和句法语义特征，在识别准确性和效率方面有显著提升。并提出结合推理知识的概念知识网络及进一步融合篇章结构的知识网络构建方法。[结果/结论]基于此知识网络模型，实现一种能够融合节点语义、拓扑结构以及类别标签信息的网络表示学习模型，能够充分挖掘并表示文本的语义及结构信息，并通过可视化和实验验证所提方法的有效性。

关键词：命名实体识别 关系提取 神经网络 表示学习 篇章结构
分类号：TP182
DOI：10.13266/j.issn.0252－3116.2021.21.019

文本信息向量化表示是自然语言处理技术的基础，如何尽可能地包含原本空间内的信息是文本向量化研究的重点。为减少计算代价，数据降维是文本表示中不可或缺的一个环节，有效的数据降维方法不仅能够减少计算量，同时有助于文本处理精度的提高，而网络表示学习作为一种图嵌入的方式将文本节点表示成低维、实值、稠密的向量形式，使文本结构信息和语义信息得到最大限度地保留，使得到的向量形式可以在向量空间中具有表示以及推理的能力，同时可轻松方便地作为机器学习模型的输入，进而可将得到的向量表示运用到常见的自然语言处理应用中。

1 相关工作

文本知识网络中的节点由各类别实体构成，边由实体之间的语法关系构成，因此文本知识网络构建首先需要进行实体抽取和实体关系识别。

1.1 实体抽取

命名实体识别技术是自然语言处理领域中重要的研究任务之一，是信息抽取、指代消歧、问答系统以及主题模型等其他任务的首要步骤。现有的命名实体抽取技术方法主要有3种，分别为基于规则和字典、基于统计机器学习以及基于深度学习。

基于规则的命名识别方法大多需要借助于词典和知识库，即根据语言学专家手工构造的规则模板，以字符串的模式正则匹配来判别文本中的实体类型。每条规则都有对应权值，当遇到规则冲突时，权值越高的规则优先级越高。基于规则的人工常用特征包括关键字、中心词或者指示词等。基于规则的命名实体识别代表性系统包括 GATE 项目中的 ANNIE 信息抽取系统，此系统依赖手工规则建立命名实体库，将每篇文章严格按照实体抽取规则定义，实现整篇文章的信息抽取。通常，当制定规则足够全面时，基于规则的命名实体抽取方法往往比其他方法的性能好。但现实中，语言现象千变万化，规则制作耗时巨大，规则之间冲突繁多，所以人工制定规则的可行性低，并且该方法需要大量的领域知识和词典为基础，系统移植性较差。

随着机器学习在自然语言处理领域的兴起，命名实体识别任务也逐渐转向基于统计的机器学习方法。

* 本文系国家重点研发计划“基于大数据的科技咨询技术与服务平台研发”（项目编号：2018YFB143502）和国家社会科学基金一般项目“数字资源知识共享与知识再利用模式与方法研究”（项目编号：21BTQ011）研究成果之一。
作者简介：刘耀（ORCID：0000－0003－3729－3866），研究员，博士，E-mail：liuy@istic.ac.cn；张越（ORCID：0000－0003－2153－6536），博士研究生；叶璐（ORCID：0000－0001－5571－9440），硕士研究生。
收稿日期：2021－06－01 修回日期：2021－09－06 本文起止页码：118－130 本文责任编辑：徐健

该方法通常将命名实体识别任务看作分类问题来处理,具有两种思路:第一种是先识别出文本中所有的命名实体的边界,然后再对实体进行分类[1]。例如 M. Collins 等[2]提出的 CoBoost 方法,训练了上下文和拼写两个分类器,再基于 AdaBoost 整合到一个适用于无监督学习的分类器。该分类器在识别人名、地名和机构名 3 类实体的准确率超过了 91%。另一种是序列化标注,即将文中每个词分配多个候选的类别标签,这些类别标签对应于命名实体所处的位置,如 IOB 系列标签格式,最后通过分类器来识别命名实体。经典的机器学习方法主要包含 HMM(Hidden Markov Model)[3]、ME(Maximum Entropy)[4]、SVM(Support Vector Machine)[5]和 CRF(Conditional Random Field)[6]等模型,在这 4 种方法中,ME 模型具有较好的通用性,但是由于需要归一化处理,计算成本比较大。而 CRF 提供了一个特征灵活、全局最优的标注框架,但是该框架收敛速度较慢,训练数据的时间较长。通常,ME 和 SVM 比 HMM 准确率要高,但是由于 HMM 采用维特比算法,所以训练速度比较快。HMM 更适用于一些对实时性有要求的应用,如短文本命名实体识别。

2011 年,R. Collobert[7]提出了采用神经网络搭建命名实体识别模型,深度学习方法不同于机器学习,无需构建繁琐的特征工程,所以一直很受欢迎。现有的神经网络模型可以根据输入颗粒度不同分为基于词汇级、基于句子级以及基于词汇和句子级的神经网络模型。在基于词汇级的模型中,将句子中每个词的词向量作为模型输入。R. Collobert 将词向量输入到 CNN + CRF 模型中,在 CoNLL2003 数据集中 F1 值达到了 89.59%。Z. Huang 等[8]在 2015 年提出了 LSTM + CRF 模型,同样使用词向量作为输入,在 CoNLL2003 数据集 F1 值达到了 85.19%。基于句子级的神经网络模型,即将整个句子表示输入到模型中,并加入句子中的位置特征来区分每一个字符。T. H. Pham 和 P. Le-Hong[9]采用将句子级表示输入到 Bi-LSTM + softmax 模型中,在越南语的命名实体识别中实现了 80.23% F1 值的效果。基于词汇和句子级的方法中,输入为词向量和单词字符卷积的组合。X. Ma 和 E. Hovy[10]将其输入到 Bi-LSTM + softmax 模型中,在 CoNLL2003 数据集中 F1 值达到了 91.21%。根据实验结果,NN(Neural Networks)和 CNN(Convolutional Neural Networks)得出的结果效果基本一致,但是加入 CRF 的句子级别效果上有明显提高。由于 Bi-LSTM 使编码结果能捕获序列信息,效果能够超过了基于丰富特征的 CRF 模型,成为目前基于深度学习的 NER 方法中最主流模型。总体来说,基于深度学习的方法无需繁杂的特征工程,即使用词向量以及字符向量就可以达到很好的效果,如果在模型中加入高质量的词典特征,性能将会进一步提高。最新的命名实体识别技术则在此基础上引入在基于神经网络的结构上加入注意力机制[11]、图神经网络、迁移学习[12]、远监督学习等技术。

1.2 关系抽取

关系抽取主要包含两个子任务:一个是检测出句子中是否包含实体对,另一个是判断实体对之间的关系。基于机器学习的关系抽取方法根据人工参与度主要分为有监督、半监督和无监督 3 种方法。有监督的机器学习方法可以分为基于特征向量的方法与基于核函数的方法,其中,基于特征向量的方法将关系抽取看作二元分类问题,使用人工标注语料获取正例和反例,通过词法分析、句法分析、语义分析得到特征集合,选取合适的分类器训练分类模型。常用的分类模型有传统机器学习模型,如条件随机场、支持向量机和最大熵分类器,还有近几年比较流行的深度学习模型,如浅层神经网络和卷积神经网络。传统的机器学习模型需要人工花费大量时间设计和选取特征,而深度学习方法采用端对端的思想,只需要预训练的词向量,人工干预较少。

半监督的机器学习方法主要采用 Bootstraping 思路解决关系抽取任务问题,即首先人工构造一批关系实例作为初始样本种子,然后利用模式训练或者模式学习的方法,总结出相应的规则,用于发现新的关系实例集合,直到得到较大规模的关系实例。DIPRE(Dual Iterative Parttern Relation Expansion)[13]系统构建了作者和书籍的关系,该系统利用少量实体关系对作为种子集合,然后获取到同时含有这两个实体的文档或者句子,将其作为标注样本,并根据标注样本建立并调整模式,最后利用该模式标注新的数据,并把新标注的数据加入到种子集合中,这样不断迭代,直到满足某种设定条件。

基于无监督的关系抽取方法大多采用模式聚类的方法。T. Hasegawa 等[14]首次提出无监督关系抽取方法,对包含命名识别实体对的文本进行聚类,把聚类集合中词频最高的词作为关系描述词,该方法在大规模的新闻领域语料上证明效果较好。M. Piasecki[15]在此基础上引入了 WordNet 语义词典来提高关系抽取模板聚类的相似度计算过程。无监督的关系抽取方法一般需要一个大规模的语料库来挖掘实体对的关系模式集

合,但是该方法难以获取置信度高的关系模式,而且难以描述关系名称。不同隐藏神经元的设置能够使 LSTM 有效避免 RNN(Recurrent Neural Network)梯度消失,从而解决了长期依赖的问题。但是单向 LSTM 只能保留历史信息,无法利用未来时刻的信息,但未来信息往往对现在时刻的信息有重要的影响。笔者采用双向长短时期神经网络(Bidirectional Long Short-term Memory,BI-LSTM),该模型包含两个 LSTM 层,分别为能够保留历史信息的前向序列和获取未来信息的后向序列,这两层 LSTM 将不同时间的信息传递给输出层。这样,Bi-LSTM 不仅能够解决长期依赖问题,还能够获取上下文信息来处理序列标注问题。

1.3 网络表示学习

网络表示学习又称图表示学习,其应用领域也相当广泛,如节点分类,节点聚类、链接预测、社区发现和推荐系统等。基于网络结构的网络表示学习主要有矩阵特征向量方法、矩阵分解方法和神经网络方法。B. Perozzi 等[16]提出 DeepWalk 算法,他们通过实验验证了节点的随机游走序列和文档中的单词一样都遵循指数定律,从而将词向量表示方法 Word2vec[17]应用在网络节点的随机游走序列中。DeepWalk 算法同样采用 skip-gram 模型对随机游走序列中每个局部窗口的节点进行概率建模,并最大化随机游走似然概率来训练模型参数[18]。这种方法只依赖于随机游走的局部信息,从而解决了矩阵特征向量方法中需要把整个邻接矩阵存储在内存中的高计算时间和空间问题。Node2vec 算法[19]同样进一步扩展了 DeepWalk 随机游走方式,通过引入两个参数 p 和 q,将深度优先和广度优先引入到随机游走策略中,从而反应了不同层面的节点之间的关系。与浅层神经网络不同的是,深层的神经网络模型能够对节点间的非线性表示进行建模。例如 SDNE(Structural Deep Network Embedding)[20]采用 laplace 矩阵对节点的一级相似度建模,然后采用无监督的深层自编码器对二级相似度建模,最终将自编码器的中间表示作为节点表示。在现实情景中,网络节点往往包含丰富的外部信息,例如在社交网络中,除了用户的好友关系,每个用户还含有丰富的文本信息,如博文等。传统的网络表示学习方法主要是刻画网络节点的拓扑结构信息,而节点的外部信息能够对拓扑结构信息进行补充,从而提高网络表示的质量。TADW(Text-Associated DeepWalk)算法[21]将文本内容特征引入到网络表示学习中,基于矩阵分解方法将关系矩阵分解为文本特征向量和两个参数矩阵,最后采用共轭梯度下降方法来更新模型参数。一些研究者还提出了 Trans 系列将知识图谱中节点间的推理结构信息考虑到网络表示学习中,例如 C. Tu 等[22]提出的 TransNet 模型通过自编码器将 TransE 关系推理模型[23]与网络拓扑结构表示相结合,在社会关系抽取任务上效果显著。网络表示学习虽然已经取得了丰富的成果,但是仍然面临巨大挑战,例如如何真正克服大规模上亿级的网络节点表示中遇到的存储、训练效率以及外部信息融合的问题等。

笔者通过科技政策文本概念关系标引,进行科技知识网络构建和网络表示学习,能为简报生成、机器翻译、问答系统等自然语言处理研究提供高质量的向量化语料信息。

2 文本知识网络构建模型

文本知识网络算法模型整体架构如图 1 所示。首先对科技政策文本进行概念与关系标引。采用 BiLSTM-CRF 深度学习模型进行概念标引,再将概念实体对的关系特征作为基于 SVM 主动学习分类器的输入,为没有标签的概念实体对进行关系标签预测。最后能够得到每篇科技政策文本中的每个句子中的概念和概念间的关系,以 json 格式存储。

然后通过概念关系分别构建科技政策知识网络以及带篇章结构的知识网络。基于知识网络模型,笔者采用网络表示学习技术对知识网络中的节点进行表示。其中知识网络中的概念表示,笔者首先采用融合节点语义、拓扑结构以及标签信息的网络表示模型,并在此基础上提出结合知识推理模型的网络表示学习模型改进方法。对于章节节点表示,笔者提出了基于 Doc2vec 的篇章节点表示方法。

2.1 概念关系标引模型

实体抽取任务的核心任务是命名实体识别,即抽取文本中所提到的人名、地名、组织、技术等。实体关系抽取即从含有实体对的句子中抽取出实体对的语义关系,关系抽取技术在自然语言理解、信息检索和知识图谱自动构建等领域具有重要的意义,能从大规模的无结构自然语言文本中抽取出结构化数据,从而提高信息处理效率。

2.1.1 基于 BiLSTM-CRF 概念标引

为了解决传统文本分析方法在科技政策文本上的局限性,BiLSTM(Bidirectional Long Short-term Memory)通过上下文特征信息可以有效地得到输出序列,但是无法表现出序列标注问题中输出标签的强依赖关系,

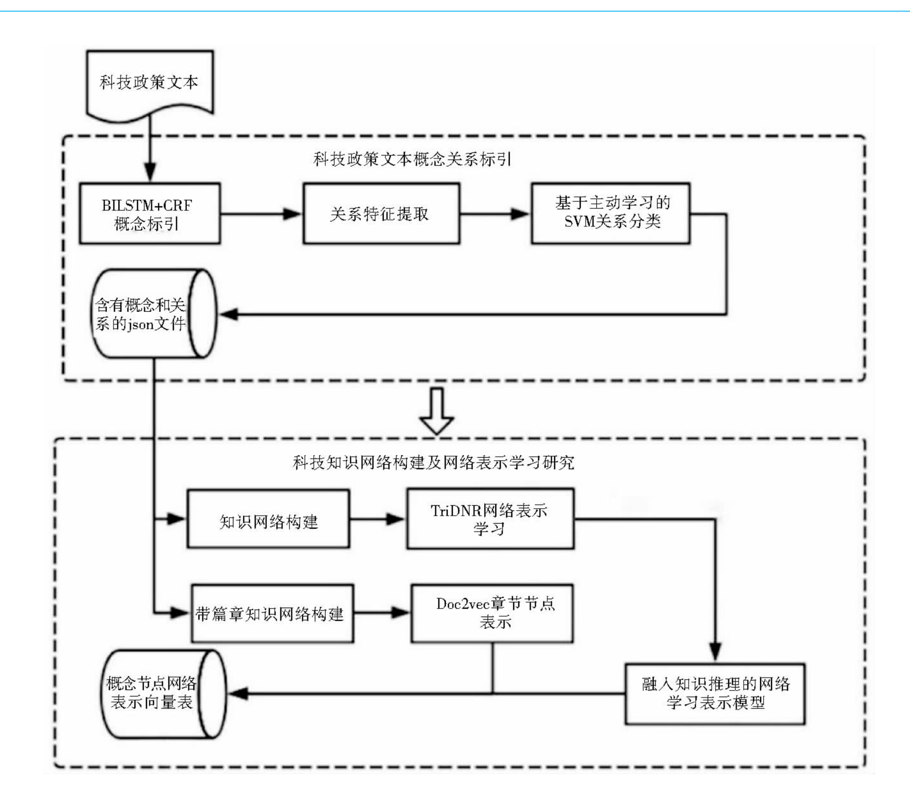


图1　文本知识网络算法框架

在本文的概念抽取任务中，在 BiLSTM 神经网络最后一层添加了用以处理序列标注的 CRF 模型，有效地解决这一问题。首先系统地对科技政策文本中的概念进行类别分析。狭义上，命名实体识别是识别出人名、地名和组织机构名这 3 类命名实体。但是在科技政策文本中，涉及的命名实体更加广泛。

首先对科技简报文本中的概念进行分类，采用概念词典、规则提取、并辅以人工标注的方法对文本中的概念进行初步标引。当词汇积累到一定程度时分析总结出以下科技简报概念词的分类体系（如果某一概念属于多个分类，则选择频次最高的分类）：①Organization：即组织名称；②Location：通常包含国家名、地名等，也可包含技术强国等概括性的术语，如“金砖国家”；③Policy：即颁发的科技政策；④Money：即政策或者技术涉及到的基金、投资以及资金；⑤Technology：通常为技术术语；⑥Field：即领域界定；⑦Energy：通常为能源类词汇；⑧Facility：即各类设备；⑨People：即人的总称或者特称；⑩System：即系统、体系或者平台；⑪Element：与其他类别词汇具有包含关系的对象；⑫Attribute：描述某一科技领域的特点；⑬Service：即国家政策提供的服务；⑭Product：即对产品的描述；⑮ Project：即提出的项目、方法或者方案。

笔者采用双向长短时期神经网络（BiLSTM），该模型包含两个 LSTM 层，分别为能够保留历史信息的前向序列和获取未来信息的后向序列，两层 LSTM 将不同时间的信息传递给输出层。这样，BiLSTM 不仅能够解决长期依赖问题，还能够获取上下文信息来处理序列标注问题。但 BiLSTM 存在的问题是无法表现出序列标注问题中输出标签的强依赖关系。笔者将中文文字以字符单元进行拆分作为 BiLSTM 神经网络模型的输入，并且采用 IOB 标注方法区分出每个句子中概念词的边界。其中，“B”标签代表概念中的第一个字，“I”标签代表概念内部的其他部分，“O”标签用于概念词汇以外的字符标示。在本文的概念抽取任务中，在 BiLSTM 神经网络最后一层添加了用以处理序列标注的 CRF 模型，有效解决了这一问题。

BiLSTM 输出的序列特征是字向量与上下文语义特征的结合，采用 Softmax 函数将隐层输出映射到标签集的概率分布向量，得到每个字符对应标签的概率分布矩阵。最后，使用 CRF 层，将概率分布矩阵在所有有效的标签序列空间中确定一个概率最高的序列路径，对应到每个字符作为最后标签。BiLSTM-CRF 模型结构见图 2。

2.1.2　基于 SVM 主动学习关系标引

关系标引的主要任务是从句子中自动抽取概念间的关系，是知识结构化的关键技术之一。笔者将关系抽取任务转化为分类任务，根据文本内容特征建立关系分类体系。首先使用词法和句法分析工具对部分语

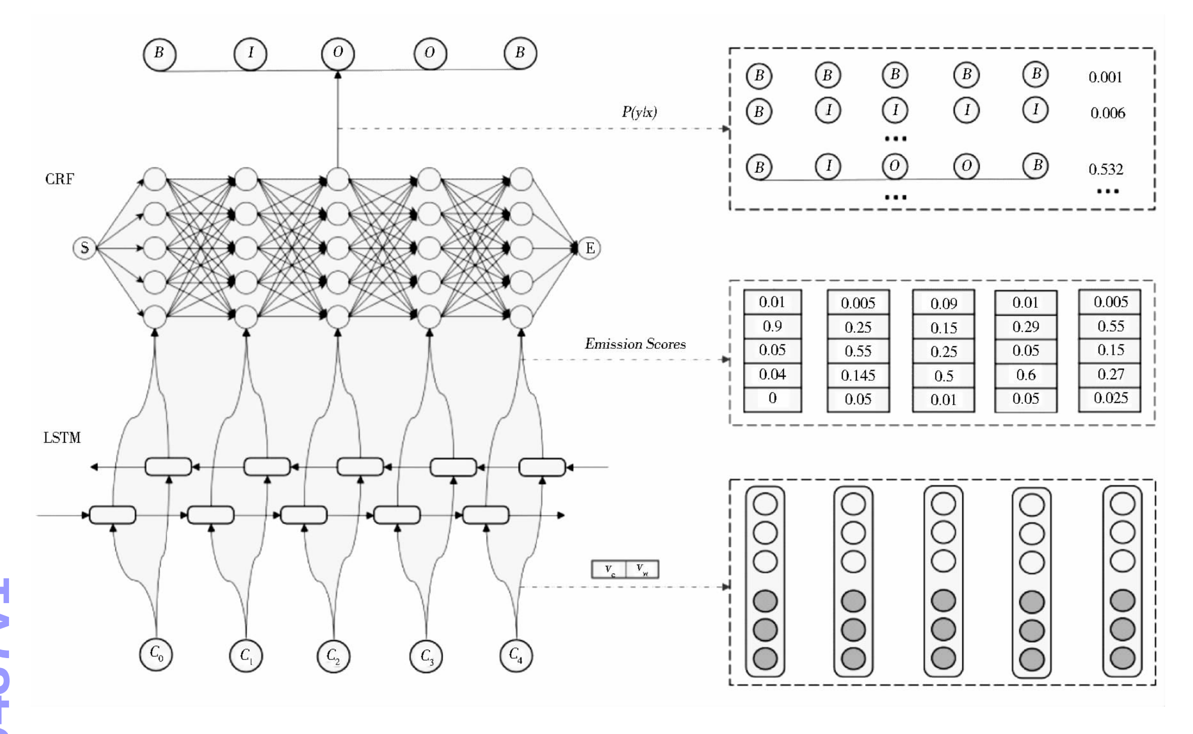


图 2 BiLSTM-CRF 模型结构

料进行处理，然后抽取出概念间的相关特征作为 SVM 分类器的输入，最后采用主动学习的方法对未标注概念关系对进行关系标引。关系标引的整体框架如图 3 所示：

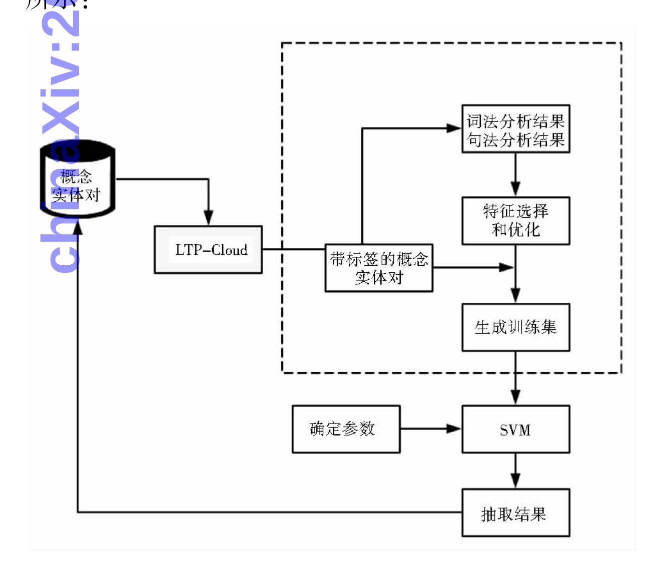


图 3 关系标引模型

笔者所研究的语料概念及其关系都具有多样性。由于概念多为名词，常常出现于主语和宾语，或者是从句的主语和宾语中，所以本研究中的关系主要是对链接概念的核心谓语进行分析，即主要分析包含“主语 + 谓语 + 宾语”或者“主语 + 谓语 + 从句（主 + 谓 + 宾）”句式结构的句子。首先对关系的种类进行预设，主要分为 5 类，分别为：①Forward：推进关系；②Mixation：融合关系；③ Backward：阻碍关系；④Inclusion：包含关系；⑤Likehood：同义关系。

笔者对关系分类采用了主动学习的方法，首先根据先验知识从候选样本中选取少量的样本进行类别标注，构造初始训练样本来训练分类器。特征主要包括基本特征和句法语义特征两类，其中基本特征主要从词法分析得出，目前研究者们已经验证了这些特征的有效性[24]。笔者选取的实体关系基本特征有：

概念类别。即为 2.1.1 中定义的 15 个概念类别，两个概念类别的结合用“ – ”字符进行连接；

概念相邻词。即获取两个概念词前面的词和后面的词，如果前后没有词，用“None”来表示；

概念间词的词性标注。即从一个概念到另一个概念中间所有词的词性标注；

两概念间的上下文环境。包括两个概念词之间的所有词。

除了基础特征，笔者还综合考虑了句法语义特征，包括依存句法分析和语义角色分析。其中，依存结构是句法分析其中重要的一方面，即通过句子中各语言单元的组成成分揭示成分之间的依存关系，其中核心谓语本身不受任何成分支配，而且能够支配其它成分的核心成分。由于概念短语是依存结构成分的一部分，那么成分之间的依存关系就能够反映出概念之间的关系。如图 4 是“纳米技术强国纷纷推进纳米技术与信息技术战略性新兴领域的融合。”的依存句法分析

刘耀，张越，叶璐.融合篇章结构的文本知识网络构建[J].图书情报工作,2021,65(21):118－130.

结构,其中“ATT”是定中关系,“ADV”是状中结构,“HED”是核心关系,“COO”是并列关系,“SBV”是主谓关系,“LAD”是左附加关系,“RAD”是右附加关系,“VOB”是动宾关系,“WP”是标点。

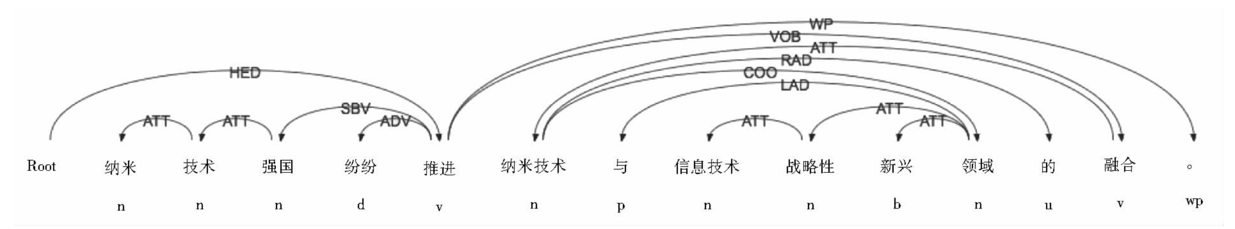


图 4　依存句法分析示

语义角色标注是一种浅层的语义分析技术,以句子的谓语为中心,分析各成分与谓语之间的关系,是自然语言理解任务中的一个重要中间步骤。例如“韩国知识经济部专门出台《纳米融合推广战略》”,其语义角色标注结果如图 5 所示。语义角色标注的结果主要包含 3 个部分,其中 A0 为施事部分,A1 为受事部分,ADV 是附加标记,语义角色标注能够在一定程度上反映了概念之间存在的语义关系。



图 5　语义角色示意

以句子“纳米技术强国纷纷推进纳米技术与信息技术战略性新兴领域的融合。”中“纳米技术强国”和“纳米技术”两个概念为例,利用 LTP 自然语言处理工具,展示本文将提取的相关特征。该句分词结果为“纳米技术强国 纷纷 推进 纳米技术 与 信息技术 战略性 新兴领域 的 融合。”特征提取的结果如表 1 所示:

表 1　关系特征示例

基础特征	概念类别	[Nation, Technology, Nation-Technology]
	概念相邻词	[None,纷纷,推进,与]
	概念间词的词性标注	[d, v]
	两概念间的上下文环境	[纷纷,推进]
依存句法分析	[3:SBV,3:VOB,10:ADV](两个概念的依存句法路径)	
语义角色标注	[“推进”A0 “推进”A1](两个概念的语义角色以及围绕的谓语词)	

支持向量机(Support Vector Machine, SVM)是一种处理分类和回归问题的机器学习算法模型,其主要思想是基于结构风险最小化原则,将训练数据集压缩到支持向量集合,从而学习到分类器。

用 SVM 实现主动学习的具体算法见表 2。

主动学习的关键是是采样策略,如何选择采样策略能够影响整个分类器的性能。为了找到合适的采样策略,在标注初期,笔者使用训练数据集中的标注数据模拟人工标注来对 LC(Least Confidence)、MS(Margin Selection)和 RS(Random Select)三种采样策略进行了性能评估。其中,LC 采样策略是选取置信度最小的 k 个样本;MS 是选择两个最高类概率但是差异性最小的 k 个样本;RS 是随机选取 k 个样本。从图 6 中可以看出,MS 的效果最好。

表 2　主动学习算法流程

主动学习算法的训练过程定义如下:
输入:初始分类器、未带标注的候选集 T、从候选样本中采样个数 n、采样策略。
1:从候选样本集 T_r 中选取 l 个样本并标注类别,来构造初始样本集 I_0,执行 $T_0 = T_r - I_0$操作。
2:进行第 i 次采样,在样本集 I_{i-1} 寻找最优分类超平面 f_i,从样本集 T_{i-1} 中最符合采样策略的 n 个样本组成集合,记作 B_i。
3:标注 B_i 样本类别。
4:执行 $I_i = I_{i-1} \cup B_i, T_i = T_r - I_i$。
5:返回 $f = f_i$
输出:分类器 f。

2.2　知识网络构建及网络表示模型

知识网络目前没有一个明确的定义,从文献[25]可知知识网络是一个集合,是指知识、信息以及知识间的关系的一类网络。其中,知识网络的节点代表知识存储单位,根据粒度不同,可以分为书刊、论文、专利、文章片段或者词。边表示知识单元之间的联系,例如在引证网络、词网络是引证关系、在共现网络是共现关系。

本文所提取的概念与概念之间的关系同样能够构成上述定义的知识网络,其中知识网络的节点是概念词汇,知识网络的边是概念词汇间的语义关系。根据以上定义,将知识网络表示为 $G = (V, E, D, L)$,其中 V

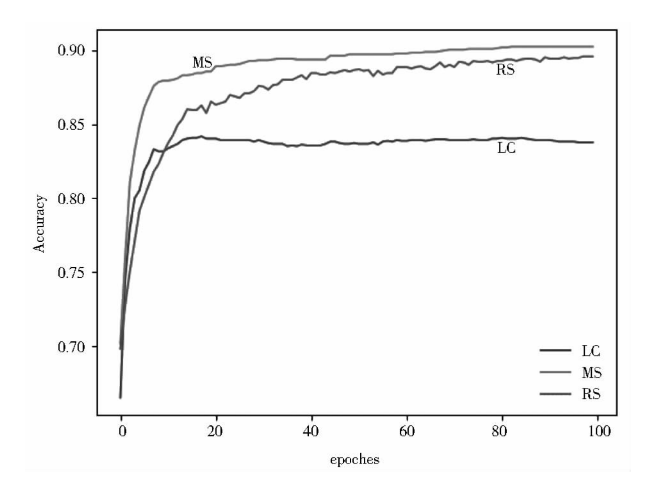


图 6　不同主动学习采样策略

$=\{v_1,v_2,\cdots,v_N\}$ 表示节点,即各个概念,$e_{i,j}=(v_i,v_j)\in$ E 表示节点之间的边,即概念之间的关系。$D=\{w_1,w_2,\cdots,w_N\}$ 表示每个节点的文本信息。$L=\{l_v,l_r\}$ 表示概念和关系标签的集合,其中 l_v 表示概念节点的类别标签,而 l_r 表示概念间关系的类别标签。

2.2.1　结合推理知识的 TriDNR 网络表示模型

网络表示学习的核心思想是学习网络中节点的低维度潜在表示。“潜在表示”的对象是网络中的节点,以及用以表示节点上下文特征、社区信息的网络拓扑结构。TriDNR 网络表示学习模型[26]通过 DeepWalk 算法获取网络节点之间的拓扑结构表示,其框架如图 7 所示, 考虑输入的网络包含节点集 V,节点 v1、v2、v3、v4、v7 各关联一个单词集 W,其中 w2、w3、w5 分别是一个长度为 2、3、5 的词序列,同时一些节点存在不同的标签属性集 C,其中 c1 为节点 v1 的类别标签,模型同时学习节点之间的关系,节点与单词的关系以及标签与单词的关系。TriDNR 模型如图 7 所示:

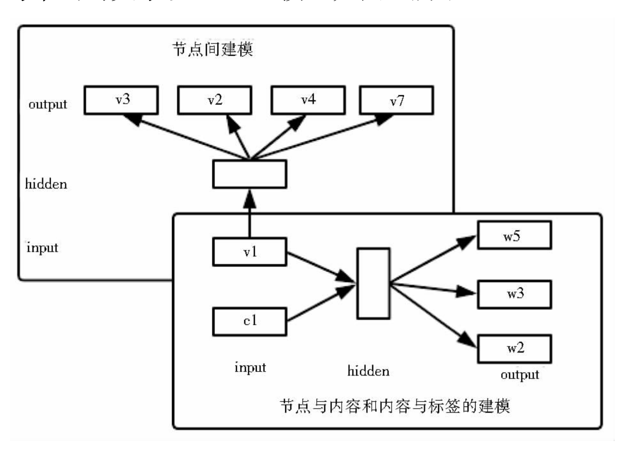


图 7　TriDNR 模型

模型由两层 skip-gram 神经网络模型组成,上层为节点的拓扑结构信息建模,下层为文本内容和文本标签建模。skip-gram 神经网络模型可以得到每个节点的表示,如同词向量的表示,Deepwalk 不同于传统的网络表示,它采用了随机游走方法,而不是邻接矩阵,解决了邻接矩阵所面临的高计算空间和时间的问题。上层结构采用 Deepwalk 算法将随机游走策略映射到每个节点表示中,该表示经过随机排序后传入下层结构。下层结构的目标函数为:

$$L=\sum_{i=1}^{|L|} log\ P\ (w_{-b}:w_b\mid c_i)+\sum_{i=1}^{|N|} log\ P\ (w_{-b}:w_b\mid v_i) \qquad 式(1)$$

从这个公式可以看出,节点内容和节点标签类似于 Doc2vec 算法,所以总体来说,笔者通过 Deepwalk 算法和 Doc2vec 算法将节点拓扑结构、节点标签和节点内容 3 个方面的信息结合起来。整体模型的目标函数是求式(2)的最大似然估计。

$$L=(1-\alpha)\ \sum_{i=1}^{N}\sum_{s\in S}\sum_{-b\leqslant j\leqslant b,j\neq o} logP(v_{i+j}\mid v_i)+\alpha\sum_{i=1}^{N}\sum_{-b\leqslant j\leqslant b} logP(w_j\mid v_i)+\ \sum_{i=1}^{|L\}}\sum_{-b\leqslant j\leqslant b} logP(w_j\mid c_i) \qquad 式(2)$$

式中,α 是平衡节点拓扑结构,节点文本内容和节点标签信息的权重,b 是窗口。其中第一个子式是计算给定一个节点,出现在这个节点周围的其他节点,可以通过 softmax 来得到,如下式:

$$P(v_{i+j}\mid v_i)=\frac{exp(v_{v_i}^{T}\ v'_{v_{i+j}})}{\sum_{v=1}^{N} exp\ (v_{v_i}^{T}\ v'_{v})} \qquad 式(3)$$

其中,v_v 和 v'_v 指的是节点 v 的输入和输出。给定节点 v,可以得到词的概率,如下式:

$$P(w_j\mid v_i)=\frac{exp(v_{v_i}^{T}\ v'_{w_j})}{\sum_{w=1}^{W} exp(v_{v_i}^{T}\ v'_{w})} \qquad 式(4)$$

同样,可以得出标签的概率,如下式:

$$P(w_j\mid c_i)=\frac{exp(v_{c_i}^{T}\ v'_{w_j})}{\sum_{w=1}^{W} exp(v_{c_i}^{T}\ v'_{w})} \qquad 式(5)$$

式(4)和式(5)共同影响节点 w_j的向量表示 v'_{w_j},而 v'_{w_j}通过反向传播影响输入 v_i,最终实现了将节点的拓扑结构、文本内容和标签三者信息共同融合的效果。

但 TriDNR 只考虑了节点之间的拓扑结构信息,没有将节点边的标签信息考虑进去。笔者借鉴 Trans 系列的知识表示学习模型中的 TransE 模型[23],将节点边的 5 种类别标签 Backward、Forward、Mixation、Likehood 和 Inclusion 融入知识网络中,这 5 种类别标签揭示了概念之间的推理关系。知识表示学习能够同时获取节点表示和边表示,通常被应用在实例链接任务中。由于本文研究只关注将关系标签映射到节点表示中,所以只将节点表示与从节点的拓扑结构,文本语义以及

刘耀，张越，叶璐.融合篇章结构的文本知识网络构建[J].图书情报工作,2021,65(21):118－130.

节点标签获得的节点表示进行向量相加求平均，最后得到的概念节点向量作为知识网络中每个概念的向量表示。

TransE 具体算法如表 3 所示：

表 3　TransE 算法

TransE 算法训练过程定义如下：
输入：训练集 $S=\{sub,rel,obj\}$，正负样本距离 γ，学习率 λ
初始化：使 *rel* 和 *sub* 与 *obj* 在 $\left(-\frac{6}{\sqrt{k}},\frac{6}{\sqrt{k}}\right)$ 均值分布。$l/\|l\|$ 正则化每一个关系
Loop：
$e \leftarrow e/\|e\|$
$S_{batch} \leftarrow \mathrm{sample}(S,b)$
$T_{batch} \leftarrow \emptyset$
for (sub, rel, obj) $\in S_{batch}$：
(sub', rel, obj') $\leftarrow$ S'(sub, rel, obj)
$T_{batch} \leftarrow T_{batch} \cup \{((sub,rel,obj),(sub',rel,obj'))\}$
end for
update embeddings $\sum_{T_{batch}} \nabla[\gamma + \|s+r-o\|_2^2 - \|s'+r-o'\|_2^2]$
end loop
输出：sub 和 obj 的节点表示 v_s 和 v_o，以及 rel 的节点表示。

2.2.2　融合篇章结构的网络表示学习模型

抽取出的概念及概念间的关系隶属于各个章节，即每个章节下都会有一个知识网络子图。当把篇章结构和知识网络结合起来，每篇科技政策文本就会形成一个上层为含有篇章结构的树状结构，底层为含有概念关系网络的知识网络模型。其中文章的题目作为 Root 节点，一级标题作为树状结构的第一层，而每个一级标题下的二级标题作为树状结构的第二层。如果有更深的层次，以此类推（通常不超过三层）。每个小标题下包含相应的概念及概念之间的关系作为底层，概念能够跨章节连接，形成网络数据结构。

概念与概念之间通过有向边相连，如同一个个句子，章节节点可以看作是一个包含了多个句子的文档。基于此分析，笔者将篇章节点分为两类：一类是在篇章树状结构底层，与概念直接相连的篇章叶子结点；另一类是其他上层篇章节点。对于叶子节点，就可以把该节点下连接的概念当作词，概念间的连线所形成的随机游走的路径当作句子，并借助于 Doc2vec 算法计算篇章节点的网络表示。对于其他上层篇章结点，采用同一层节点求和取平均数的方法，层层计算，直到取得根节点的向量表示。

对于篇章叶子节点表示所采用的 Doc2vec 方法的模型如图 8 所示。其中，w 是概念节点，v 是概念节点的网络表示，paragraph matrix 在本模型中代表章节点的表示。然后章节节点与概念向量进行连接或者简单相加来对下一个概念进行预测，从而构建浅层神经网络模型，最后通过训练模型就能够得到篇章叶子节点的表示。这样不同的章节就可以根据自己相连的不同概念获得不同的向量表示，但是不同章节中相同的概念具有相同的概念表示。由于 Doc2vec 算法是一种无监督的算法，即可以对没有标注的数据进行训练，所以此模型能快速高效地得到章节节点的向量表示。

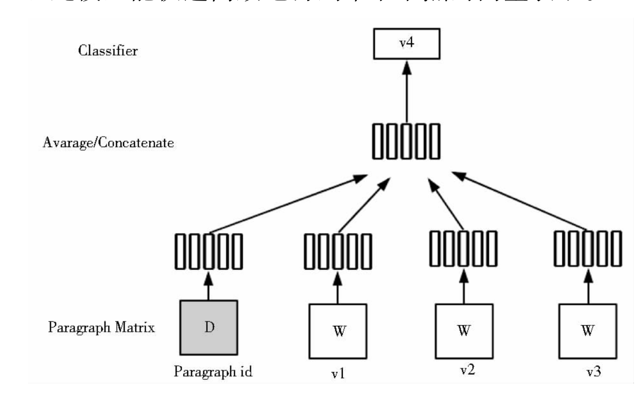


图 8　篇章叶子节点表示方法模型

3　实验设置与结果分析

3.1　概念关系标引实验

对于概念识别，笔者构建了以词向量作为输入的 BiLSTM + CRF 命名实体识别深度学习模型。对于关系识别，为实现从高层次语义挖掘概念间的语义知识，笔者设计了概念实体对之间关系的概念类别、概念相邻词、概念间词的词性标注、两概念间的上下文环境 4 个基本特征和依存句法分析、语义角色分析两个语义特征，并训练 SVM 分类器证明本实验选取特征的有效性。

3.1.1　BiLSTM-CRF 概念提取实验

实验数据集如下：

BiLSTM-CRF 概念提取实验使用的数据集是科技部公开发布的近 20 年的科技参考，共 1 000 篇。按照不同的年份进行混合，人工将数据集分为 10 份，使用规则并辅以人工对概念进行标注其中的一份，共 4 340 个句子，4 790 个不重复概念，采用 BIO 标注每个字后，有将近 23 万个字符标签，将其以 8：1：1 分为训练语料、测试语料、验证语料。

实验结果与分析如下：

笔者将概念抽取问题转化为序列标注问题，为了验证本文使用方法的有效性，实验中对比了传统的 CRF 方法、BiLSTM 方法以及 BiLSTM-CRF 3 种使用较多的概念抽取方法，由于涉及的概念类别较多，共有

15 个类，所以最终结果对所有类的结果取平均数。结果如表 4 所示：

表 4　概念抽取结果对比

模型	Precision	Recall	F1
TextRank + 句式	32.34	31.05	31.68
CRF	55.67	50.37	53.42
BiLSTM	63.97	52.02	57.38
BiLSTM + CRF	70.89	63.59	67.32

由表 4 可以看出，BiLSTM 概念抽取方法的性能要优于传统的机器学习方法，而在 BiLSTM 模型中加入 CRF 层能够提升概念抽取的效果。对各个类别识别结果进行分析，Organization 概念类别识别的效果最好，能达到 80.25% 的准确率。但是，其他如 Attribute 或者 Service 由于在训练数据集中所占比例较少，最终识别效果较差。

3.1.2　SVM 概念关系识别实验

实验数据集如下：

SVM 概念关系实验人工标注了 100 篇概念关系（分为推进关系、融合关系、阻碍关系、包含关系和无关系，共 5 种关系），共 1 980 条关系作为 SVM 的初始训练样本，其中的 80% 作为训练语料，剩下的 20% 作为测试语料，未标注的候选集是从剩下 900 篇的未标注概念中产生，关系数约为 22 500 条关系。笔者采用主动学习的方法，每次抽取 200 个对分类器性能影响最大的样本进行分类预测，校对后分类正确的集合将被加入到训练集中进行再次训练。概念关系标引结果如图 9 所示：


```
{
    "sentence": "此外，将建立各领域的\"纳米融合创新联盟\"，促进现有支柱产业与纳米技术的融合及产业化。",
    "concept": [
        {
            "id": "T93",
            "word": "纳米技术",
            "type": "Technology",
            "start": "1677",
            "end": "1681"
        },
        {
            "id": "T94",
            "word": "产业化",
            "type": "Attribute",
            "start": "1685",
            "end": "1688"
        },
        {
            "id": "T187",
            "word": "纳米融合创新联盟",
            "type": "Organization",
            "start": "1658",
            "end": "1666"
        }
    ],
    "relationship": [
        {
            "id": "R42",
            "word1": "T187",
            "word2": "T94",
            "relationtype": "Forward"
        }
    ]
},
```


图 9　概念关系标引结果

实验结果与分析如下：

表 5 是针对两组不同特征，SVM 分类器在不同关系类别上的分类效果。其中，第一组实验选取基本特征，第二组实验选用了基础特征和句法语义特征（依存句法分析和语义角色标注）。基本特征提取的是命名实体间的词间关系，缺少句级的语法特征信息，但句内关系具有较强的组织关联性，对挖掘深层语义信息具有辅助意义。从实验结果也可以看出，第二组实验使用句法语义特征后，实体关系抽取的效果优于第一组实验，证明了使用句法语义特征的有效性。

表 5　SVM 关系抽实验结果统计

特征类型	分类类别	Precision	Recall	F1
基本特征	推进关系	60.49	47.89	53.31
	融合关系	67.55	63.29	65.42
	阻碍关系	55.67	48.67	52.10
	包含关系	66.48	63.66	65.04
	无关系	71.38	73.21	72.28
整体		64.31	59.34	61.83
基本特征 + 句法语义特征	推进关系	63.16	53.98	58.21
	融合关系	73.16	70.89	72.01
	阻碍关系	63.12	66.53	64.78
	包含关系	70.19	67.54	68.84
	无关系	74.03	78.61	76.25
整体		68.73	67.51	68.02

刘耀，张越，叶璐.融合篇章结构的文本知识网络构建[J].图书情报工作,2021,65(21):118－130.

3.2 知识网络构建及网络表示实验

笔者对 TriDNR 网络表示学习模型进行改进,采用能够融合网络节点拓扑结构、节点内容、节点标签以及节点之间推理信息 4 个方面的信息的网络表示学习模型,从不同方面弥补节点表示不全面的问题。对于融合篇章结构的知识网络,笔者在简单向量相加的方法的基础上进行改进,采用 Doc2vec 算法来表示篇章节点,最后通过 TextRank 算法对篇章节点重要性进行排序,从而验证利用笔者提出的方法生成的篇章节点向量表示的有效性。

3.2.1 结合推理知识的 TriDNR 网络表示学习实验

实验数据集如下:

得到 BiLSTM-CRF 训练模型之后,按次序从其他 9 份中选 1 份进行概念预测,之后进行人工校对,最后将 1 000 篇中将近 45 000 个的句子进行概念标引,将其中具有关系的实体对作为语料,共约 35 000 个不重复概念,约近 28 000 多条关系,融合成大概念网络。

实验结果与分析如下:

为了验证本文所采用方法的有效性,实验中将其他概念节点表示方法与本文使用方作对比。为了公平起见,笔者将不同方法的表示维度都设置为 300,由于想要每个概念节点涵盖的信息不仅仅局限于之间近邻的节点,所以使控制深度优先搜索的参数 p 大于 1。表 6 是不同方法之间的对比结果,选取了节点表示一些常用的方法,其中 DeepWalk、Node2ve 只是考虑节点的拓扑结构,Doc2vec 只考虑节点的文本信息和标签信息,DW + Doc2vec 是将拓扑结构信息、文本内容信息和标签信息的表示上相加, TriDNR 是将三者信息共同训练,笔者提出的方法则是将节点的知识推理信息融合到节点表示中。

表 6　各类概念节点表示方法对比实验结果

% p	DeepWalk	Doc2vec	DW + Doc2vec	Node2vec	TriDNR	本文方法
10	0.243	0.275	0.341	0.312	0.476	0.512
30	0.315	0.361	0.407	0.386	0.553	0.593
50	0.394	0.428	0.478	0.443	0.618	0.649
70	0.431	0.452	0.501	0.492	0.642	0.683

通过实验结果可以得出,只考虑概念节点的拓扑结果或者节点文本内容的 F1 值最低,尤其在训练集较少的情况下,当将二者结合后,分类效果有较高的提升。从结合方法来看,通过模型训练的方法比直接将两部分信息相加更有效。而笔者提出的方法增加了知识推理模型,所以效果较前几种都有所提升。

利用网络表示学习技术将知识网络的概念节点映射成了一个 300 维的向量表,通过该向量表,就能够将知识网络的拓扑结构信息、语义信息、节点标签信息以及节点之间的推理信息都融入到节点表示中,最终被应用在其他机器学习或深度学习模型中,其可视化效果如图 10 所示:

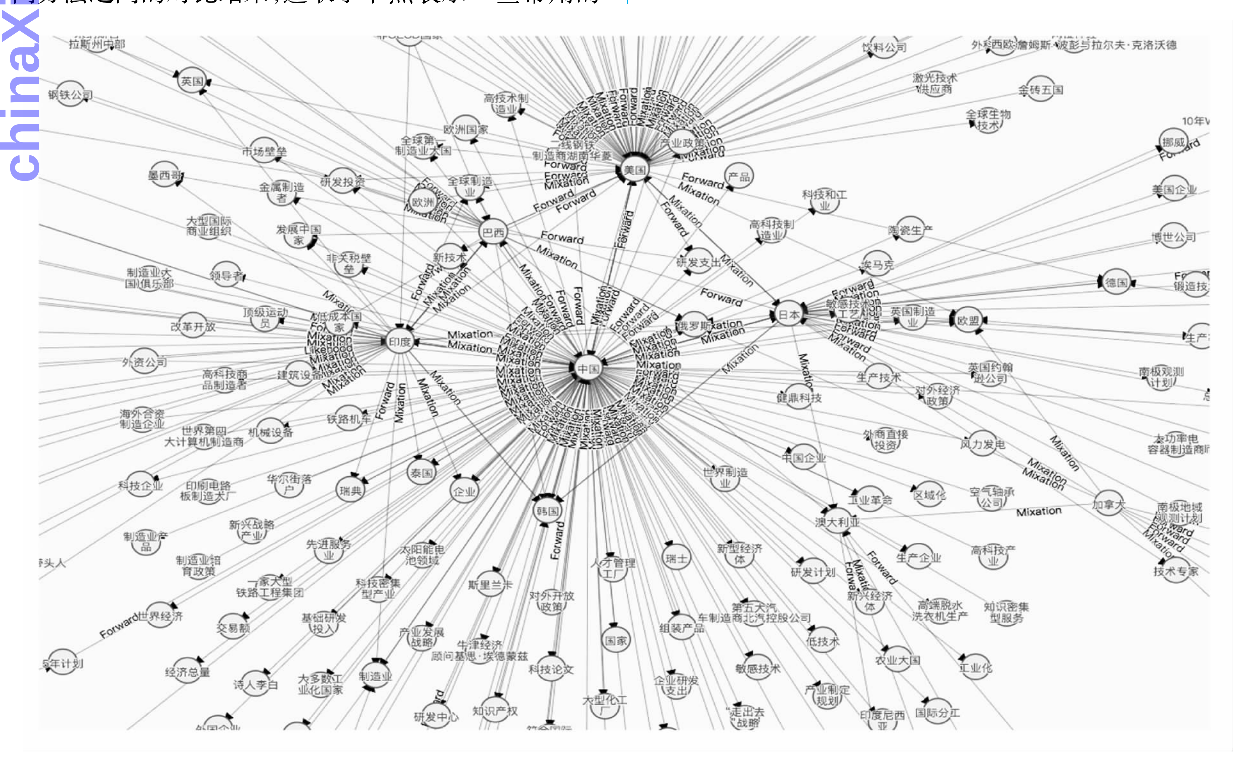


图 10　结合推理信息的 TriDNR 可视化

3.2.2 融合篇章结构的网络表示学习实验

实验数据集如下：

在文本知识网络构建实验中，以“德国工业 4.0”为主题，挑选了 10 篇相关文本，将章节节点的表示作为输入，通过实验中对章节的排序结果与原文的章节的排序进行比较，从而证明本文方法的有效性。

实验结果与分析如下：

图 11 是融合篇章结构的知识网络示例图，从图中可以明显地看到树状篇章将知识网络组织起来。不同章节下的相同的概念进行合并，不同概念相互连接，最终形成树状结构与网络结构相结合的融合篇章结构的知识网络模型。

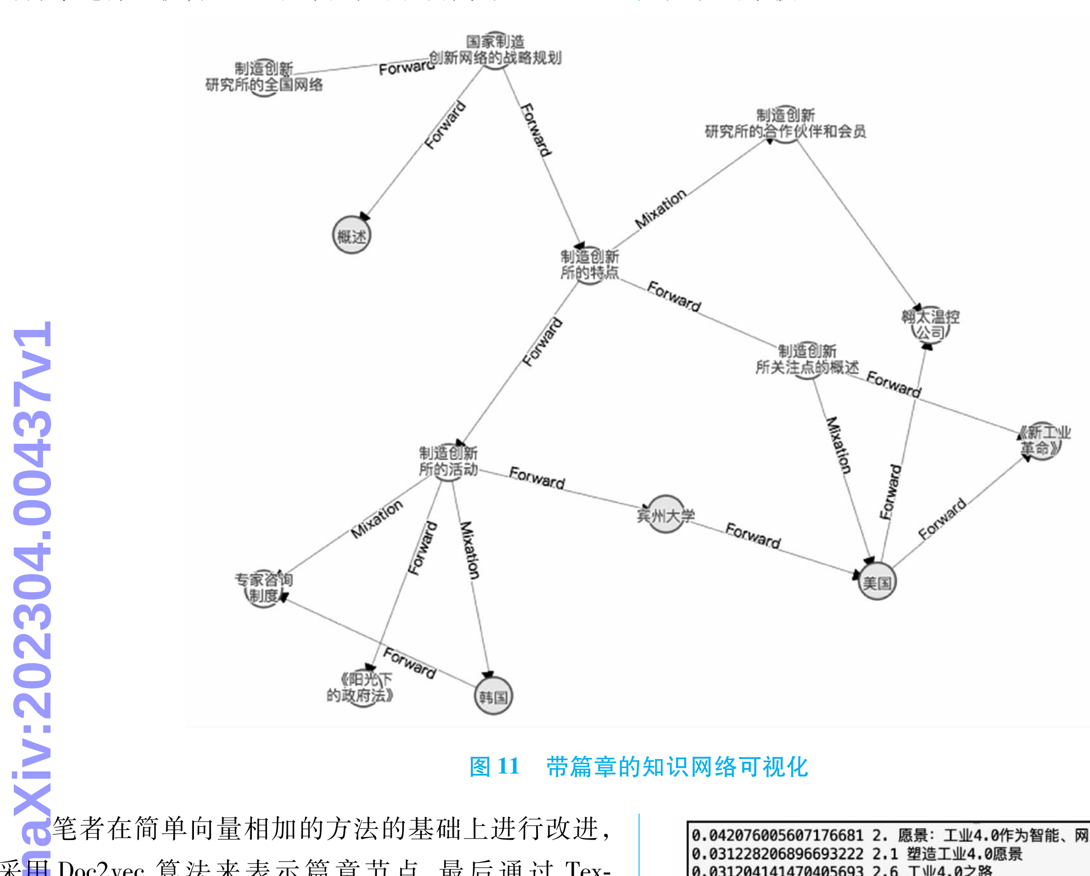


图 11　带篇章的知识网络可视化

笔者在简单向量相加的方法的基础上进行改进，采用 Doc2vec 算法来表示篇章节点，最后通过 TextRank 算法对篇章节点重要性进行排序。以德国发布的《实施“工业 4.0”攻略的建议》为例，实验输出了前 8 个排序最高的章节节点，图 12(a)是只用了 TextRank 算法的排序结果，图 12(b)是采用了概念节点平均值求和的方法得到章节节点表示，图 12(c)是通过本文方法来表示章节节点的排序结果。

融合篇章结构的文本知识网络的有效性可通过单篇文本篇章的重要性排序进行证明，而篇章的重要性可以由两个指标来确定：一个是篇章中包含的内容，即包含的概念越丰富，那么这个章节就越重要；另一个是篇章所处的位置，通常认为题目比一级目录重要，二级目录比三级目录重要，也就是说在树状结构中，相较于下层篇章节点，上层篇章节点更具概括性，包含的重要概念更多，所处的位置更加重要。从实验结果可以发现，使用本文方法得到的章节节点排序结果，前 3 个结果均为一级标题，优于其他两种方法，而采用概念节点

```
0.042076005607176681 2. 愿景：工业4.0作为智能、网络化世界的一部分
0.031228206896693222 2.1 塑造工业4.0愿景
0.031204141470405693 2.6 工业4.0之路
0.030919204916247241 3. 双重战略：成为领先的市场和供应商
0.030919204916247241 5.4 安保是工业4.0成功至关重要的因素
0.030652227184829544 5.3 为工业提供一个全面宽敞的基础设施
0.030652227184829543 3.3.3 纵向集成和网络化制造系统
0.030519294657362788 2.2 在工业4.0下未来会是什么样子
```

(a) TextRank 节点排序结果

```
0.053385834857927576 2. 愿景：工业4.0作为智能、网络化世界的一部分
0.04927265931881513 3. 双重战略：成为领先的市场和供应商
0.04701077405687689 2.1 塑造工业4.0愿景
0.04573282513747128 5. 优先行动领域
0.04426482468843093 5.5 数字化工业时代工作的组织和设计
0.04374719956100772 2.6 工业4.0之路
0.04302327508521478 3.3.3 纵向集成和网络化制造系统
0.04140968456483198 5.4 安保是工业4.0成功至关重要的因素
```

(b) 概念向量相加排序结果

```
0.053179667466278425 2. 愿景：工业4.0作为智能、网络化世界的一部分
0.04240042462095923 3. 双重战略：成为领先的市场和供应商
0.042104344793404644 5. 优先行动领域
0.0413583841679672 2.1 塑造工业4.0愿景
0.041164427328138996 5.4 安保是工业4.0成功至关重要的因素
0.040801793464504135 5.3 为工业提供一个全面宽敞的基础设施
0.040317800639493085 2.6 工业4.0之路
0.040177119269556284 3.3.3 纵向集成和网络化制造系统
```

(c) 本文方法排序结果

图 12　不同方法章节节点重要性排序

平均值求和的方法优于只用 TextRank 的方法。本文

使用的方法没有把第四章和第六章的节点排名靠前，原因是这两个章节下面没有二级标题,篇幅较短,所以能够提取的概念并不多。

4 结语

笔者以研究科技政策文本为研究对象,通过概念与关系标引技术和融合推理知识和篇章结构的知识网络构建,实现对文本与向量空间的映射,完成了语义信息的深度挖掘,为自然语言处理的各项应用提供丰富的结构和语义信息。笔者采用 BiLSTM + CRF 深度学习模型进行概念标引,并分析了概念实体对之间关系特征来训练 SVM 分类器。实验发现,深层句法分析特征(依存句法和语义角色)能够提高部分关系类别识别的准确率。笔者还采用了主动学习的方法,能够有效地将科技政策文本中的概念提取出来并为概念实体对标引关系,大大降低了人工标引的工作量。

在科技政策文本的概念与关系标引的基础上,笔者首先采用了融合节点语义、拓扑结构以及标签信息的网络表示模型,并在此基础上开创性地提出结合知识推理模型的网络表示学习模型改进方法和带篇章结构的知识网络模型的篇章节点表示,通过可视化和重要节点排序实验验证了笔者提出的文本网络表示方法的有效性,生成的向量网络可有效服务于文本挖掘、信息检索、问答系统等自然语言处理任务。

参考文献:

[1] 张晓艳, 王挺, 陈火旺. 命名实体识别研究[J]. 计算机科学, 2005, 32(4):44－48.

[2] COLLINS M, SINGER Y. Unsupervised models for named entity classification[C]//1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora. Stroudsburg: ACL, 1999.

[3] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name[J]. Machine learning, 1999, 34(1－3): 211－231.

[4] CURRAN J R, CLARK S. Language independent NER using a maximum entropy tagger[C]//Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003. Stroudsburg: ACL, 2003: 164－167.

[5] MCNAMEE P, MAYFIELD J. Entity extraction without language-specific resources[C]//Proceedings of Association for Computational Linguistics. Stroudsburg: ACL, 2002: 1－4.

[6] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//Association for computational linguistics. Stroudsburg: ACL, 2003: 188－191.

[7] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(Aug): 2493－2537.

[8] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer science, 2015:1－10. [2021－08－27]. https://arxiv.org/pdf/1508.0－1991.pdf.

[9] PHAM T H, LE-HONG P. End-to-end recurrent neural network models for vietnamese named entity recognition: Word-level vs. character-level[C]//International conference of the pacific association for computational linguistics. Singapore: Springer, 2017: 219－232.

[10] MA X, HOVY E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint, 2016, arXiv:1603.01354.

[11] WANG W, CHANG L, BIN C, et al. ESN-NER: entity storage network using attention mechanism for chinese NER[C]//Information processing and cloud computing. New York: ACM, 2019: 1－8.

[12] 余传明, 黄婷婷, 林虹君, 等. 基于标签迁移和深度学习的跨语言实体抽取研究[J]. 现代情报, 2020, 40(12):3－16,35.

[13] BRIN S. Extracting patterns and relations from the world wide web[C]// International Workshop on the World Wide Web and databases. Berlin: Springer, 1998: 172－183.

[14] HASEGAWA T, SEKINE S, GRISHMAN R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Stroudsburg:ACL, 2004: 415.

[15] PIASECKI M, RAMOCKI R, KALINSKI M. Information spreading in expanding wordnet hypernymy structure[C]//Proceedings of the international conference recent advances in natural language processing. New York: ACM, 2013: 553－561.

[16] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2014: 701－710.

[17] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. MA: MIT Press, 2013: 3111－3119.

[18] 涂存超, 杨成, 刘知远, 等. 网络表示学习综述[J]. 中国科学:信息科学, 2017(8):32－48.

[19] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[C]//Proceedings of the 22th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2016: 855－864.

[20] WANG D, CUI P, ZHU W. Structural deep network embedding[C]//Proceedings of the 22th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2016: 1225－1234.

[21] YANG C, LIU Z, ZHAO D, et al. Network representation learning

with rich text information[C]//International joint conference on knowleclge discovery and data mining. New York: ACM, 2015: 2111 - 2117.
[22] TU C, ZHANG Z, LIU Z, et al. TransNet: translation-based network representation learning for social relation extraction[C]//IJCAI. New York: ACM, 2017: 2864 - 2870.
[23] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]//Advances in neural information processing systems. New York: ACM, 2013: 2787 - 2795.
[24] 刘丹丹, 彭成, 钱龙华, 等. 词汇语义信息对中文实体关系抽取影响的比较[J]. 计算机应用, 2012, 32(8):2238 - 2244.
[25] 刘向, 马费成, 陈潇俊, 等. 知识网络的结构与演化——概念与理论进展[J]. 情报科学, 2011(6):801 - 809.
[26] PAN S, JIA W, ZHU X, et al. Tri-party deep network representation[C]// International joint conference on Artificial Intelligence. New York: ACM,2016: 1895 - 1901.

作者贡献说明:
刘耀:提出研究思路与研究方向;
张越:设计研究方案,进行实验,论文起草;
叶璐:负责数据准备,论文修订。

Construction of Text Knowledge Network Integrating Discourse Structure

Liu Yao[1] Zhang Yue[2] Ye Lu[3]
[1] Institute of Scientific and Technical Information of China (ISTIC), Beijing 100038
[2] Michigan State University, East Lansing 489132
[3] School of Software & Microelectronics, Peking University, Beijing 100871

Abstract: [**Purpose/significance**] Text vectorization is a necessary pre-processing process in the fields of text mining, information retrieval, sentiment analysis, etc. It is an urgent problem to make node vectors contain rich and effective semantic and structural information. [**Method/process**] At first, this paper analyzed the text characteristic of science and technology policy. According to the classification system of the concept and the relationship between the concepts, this paper used BiLSTM - CRF algorithm and SVM respectively to extract index the concepts and their relations automatically. Meanwhile, the model integrated basic characteristics and syntactic semantic features in feature engineering, leading to a boost in recognition accuracy and efficiency. This article also put forward the concept knowledge network combining reasoning knowledge and the knowledge network construction method of furtherly integrating discourse structure. [**Result/conclusion**] Based on this knowledge network model, this paper implements a network representation learning model that can integrate node semantics, topology structure and category label information. It can fully exploit and represent text semantic and structural information, and through the visualization and experiment to verify the effectiveness of the proposed method.

Keywords: named entity recognition relationship extraction neural network representation learning discourse structure